Measuring Progress in Second Language Pronunciation Learning using Automated Assessment Metrics

Anonymous ACL submission

Abstract

A teaching strategy using repetition has been popular for second language (L2) pronunciation learning. Built upon the strategy, the effectiveness of repetition is known to be enhanced by feedback. This study investigates the effectiveness of repetition with and without feedback as pronunciation learning strategies for Chinese learners of English, utilising multiple automated pronunciation assessment metrics. The use of automatic pronunciation assessment helps avoid the subjectivity of human evaluation, which often shows weak correlations among raters, making automated methods more reliable. A novel corpus, Repetition-based Pronunciation Improvement (RPI), was collected from 50 Chinese learners divided into two groups: repetition only (RPI_G1) and repetition with feedback (RPI_G2). Eighteen pronunciation assessment metrics, including automatic phone recognition, self-supervised models, and Goodness of Pronunciation (GOP) were used to evaluate learner pronunciations over 12 repetitions of 7 pseudo-words. Results show RPI_G2 demonstrated positive learning rates across most metrics, while RPI_G1 showed negative learning rates, indicating the importance of feedback for pronunciation improvement. Analysis of the metrics revealed varying levels of consistency and correlation, with self-supervised models showing high correlation.

1 Introduction

The mastery of English pronunciation is crucial for learners of English as a second language (L2). Accurate pronunciation is essential for clear communication, boosting confidence, and enhancing cultural understanding in L2. Each learner brings unique qualities and behaviours to their learning journey, creating a diverse landscape of approaches to pronunciation improvement (Gilakjani and Ahmadi, 2011). One effective learning strategy for pronunciation learning is an exercise focusing on pronunciation of words involving minimal acoustically confusable pairs. This strategy has been shown to enhance pronunciation skills in L2 learners (Darcy, 2018; Gilakjani, 2012). Repetition of words is another strategy. It allows learners to intentionally practice saying words and sounds repeatedly to strengthen and build confidence in their pronunciation (Larsen-Freeman, 2012). When combined with corrective feedback, repetition is enables learners to not only practice and identify errors independently but also receive guidance on how to improve their pronunciation (Saeli et al., 2021). Despite the existing literature emphasising the importance of integrating various L2 pronunciation learning strategies, incorporating automated assessment metrics, and considering the specific characteristics of L2 learners, significant gaps remain in the field of Computer-Assisted Pronunciation Training (CAPT). Luo (2016) and Tejedor-García et al. (2020) identified a lack of standardised guidelines for evaluating pronunciation improvement in CAPT and a shortage of objective studies on the effectiveness of Automatic Speech Recognition (ASR) and Text-to-Speech (TTS) systems within CAPT. Furthermore, recent studies (Kunihara et al., 2022; Malucha, 2022) used a limited number of evaluation matrices, suggesting a need for exploring alternative L2 pronunciation learning strategies. To address these shortcomings, this study shows the effectiveness of repetition with and without feedback by utilising multiple automated pronunciation assessment metrics for L2 learners. The effectiveness is investigated with a novel corpus, the Repetition-based Pronunciation Improvement (RPI) corpus, which was collected for this research. This corpus focuses on Chinese learners of English, where the demand for effective learning strategies is high. In addition to formal L2 pronunciation assessment metrics, this study builds on recent successes in utilising self-supervised learning models (Kim et al., 2022; Islam et al., 2023), such as Wav2Vec 2.0, Hubert models, WavLM models, and XLS-R. The exploration of these models for pronunciation assessment opens up their potential usefulness for capturing pronunciation development and complements traditional assessment methods. The experiments are designed to answer the following research questions:

Research Question 1 (RQ1): How is the effectiveness of L2 pronunciation learning strategies using repetition influenced by providing feedback?

Research Question 2 (RQ2): How useful are different pronunciation assessment metrics for evaluating L2 pronunciation learning?

To address these questions, data was collected specifically examining the effectiveness of repetition and repetition with feedback as L2 pronunciation learning strategies. The study also incorporates a comparison of various pronunciation assessment metrics, enriching the understanding of the nuances in pronunciation assessment. Through these multifaceted investigations, this research aims to contribute to the analysis of L2 learning and offer insights for L2 teachers.

2 Pronunciation assessment

Pronunciation errors can be categorised into two main types: phonetic (segmental) errors and prosodic errors (Chang, 2021). Phonetic errors involve the mispronunciation of individual sounds, such as vowels and consonants, and can manifest as insertion, deletion, or substitution errors. In contrast, prosodic errors pertain to broader elements influencing the pronunciation of entire words or sentences, including stress, rhythm, and intonation (Islam, 2020). Chinese L2 English learners encounter various challenges in pronunciation, with research indicating that they experience difficulties in both segmental and prosodic aspects (Han, 2013). Several studies have examined the influence of first language (L1) backgrounds on the perception and production of L2 (Zhang and Xiao, 2014; Richards, 2011). For example, the 'th' sounds (/θ/ and /ð/) in words like "think" and "this" are absent in Chinese, leading to common substitutions with /s/, /z/, /t/, or /d/. The English 'r' and 'l' sounds also pose difficulties, as Chinese learners often merge them into a single sound. Furthermore, the distinction between /v/ and /w/ is non-existent in Chinese, causing confusion between words such as "vine" and "wine". Feedback plays a crucial role in pronunciation learning, as it helps learners identify and correct their errors. Saito and Lyster (2012) found that corrective feedback, particularly explicit correction and metalinguistic explanation, led to significant improvements in the pronunciation of Japanese learners of English. Similarly, Lee (2013) demonstrated the effectiveness of corrective feedback in improving the pronunciation of Korean learners of English, highlighting the importance of immediate and explicit feedback. However, these studies relied on human evaluators, which can be subjective and time-consuming. Automated L2 pronunciation assessment offers objective evaluations based on predefined criteria, with the added benefit of potentially eliminating subjective biases. Advancements in recent years have significantly improved the field of pronunciation assessment and its utilisation in CAPT (Agarwal and Chakraborty, 2019; Rogerson-Revell, 2021; Korzekwa et al., 2022). Different automatic pronunciation assessments can be employed for each type of pronunciation error (Kheir et al., 2023). Using automatic phone recognition in L2 pronunciation assessment allows the processing of a learner's spoken input. The audio is transformed into streams of features, which then undergo recognition with implicit phonetic segmentation. Individual phonemes are identified and compared to a native speaker-based reference model (Yeo et al., 2023; Li et al., 2020). The L2 learner's phoneme accuracy is evaluated by computing the Phoneme Error Rate (PER), which is the ratio of the total number of phoneme errors, including inserted, deleted, and changed phonemes, to the overall number of phonemes in the reference. Inspired by the recent achievements of self-supervised learning models in speech-related tasks, including speech recognition, emotion recognition, speaker verification, and language identification, as demonstrated in prior works (Ravanelli et al., 2020; Morais et al., 2022; Chen et al., 2021; Tjandra et al., 2022), the L2 pronunciation assessment field also incorporates self-supervised learning models (Kim et al., 2022; Islam et al., 2023). Goodness Of Pronunciation (GOP), initially introduced by Kim et al. (1997), is a likelihood-based mispronunciation detection algorithm based on Hidden Markov Model-Gaussian Mixture Model (HMM-GMM) Automatic Speech Recognition (ASR) models. It provides phoneme scores and thus allows the detection of errors at the phoneme level. Building upon that, Zhang et al. (2008) proposed enhancements of GOP aimed at refining the GOP scoring methodology to improve its effectiveness. These have been shown to outperform previous methods on phoneme and utterance-level assessment tasks (Sheoran et al., 2023; Kanters et al., 2009). Gong et al. (2022) introduced a GOP feature-based Transformer (GOPT), which integrates with various acoustic models. The authors report a Pearson correlation coefficient (PCC) of 0.612 with human expert evaluations on the speechocean762 corpus at the phone level (Zhang et al., 2021). This demonstrates the potential of transformer-based models in capturing the nuances of pronunciation assessment. Despite the advancements in automated pronunciation assessment, several limitations persist in existing studies. Many studies have focused on a single metric or a limited set of metrics, making it difficult to compare the effectiveness of different approaches (Hu et al., 2015). Furthermore, the lack of large-scale, publicly available corpora with detailed annotations for pronunciation assessment hinders the development and evaluation of new methods (Wang et al., 2018; Zhang et al., 2021). The computation of automatic assessment metrics relies on the availability of a substantial amount of training data that is directly relevant to the specific task. However, obtaining such data can be challenging and costly. Some available corpora have limited public accessibility, and among these, only a few contain detailed transcriptions. Even fewer provide manual assessments of prosodic features, fluency, and overall proficiency scores. Several corpora featuring L2 learners speaking English have been developed to address these challenges . One such example is the ISLE corpus, which includes recordings of 23 intermediate-level speakers each for German and Italian-accented English (Menzel et al., 2000). Speechocean762 is a dataset specifically designed for pronunciation assessment, featuring a total of 5,000 English utterances from 250 Chinese speakers. Each utterance in the dataset is assessed by five experts at the utterance, word, and phoneme levels (Zhang et al., 2021). To address the limitations of existing studies and explore the effectiveness of feedback and repetition in L2 pronunciation learning, this study utilises multiple automated pronunciation assessment metrics and collects a novel corpus focusing on Chinese learners of English. By comparing the performance of learners who receive feedback during repetition with those who do not, this study aims to provide insights into the role of feedback in pronunciation improvement. Additionally, by evaluating the consistency and correlation between various assessment metrics, this study seeks to identify the most useful metrics for evaluating L2 pronunciation learning.

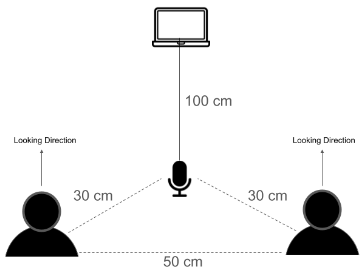

Figure 1: Recording setup for RPI_G2 group.

3 Repetition-Based Pronunciation Improvement Corpus

To measure progress in L2 pronunciation learning using repetition as a learning strategy, a new corpus was collected, as specific data in sufficient amounts was not readily available. This section describes the participants, data collection process, and corpus details for the Repetition-Based Pronunciation Improvement (RPI) corpora. A total of 50 L2 learners, who are Chinese native speakers and students at the university, participated in the RPI corpora. Among them, 43 were within the age group of 20-30, and seven were within the age group of 31-40.

3.1 Words List

The word list consisted of seven pseudo-words, each comprising 6-7 phonemes. Notably, existing literature (Khanal et al., 2021; Wang and Chen, 2020; Chan, 2007) has identified two to three of these phonemes as challenging for Mandarin speakers learning English. Pseudo-words or nonce words are terms used in linguistics to describe words created for a specific purpose and do not have an established meaning in the language (Keuleers and Brysbaert, 2011). The use of pseudo-words aimed to provide a more authentic evaluation of learners' ability to reproduce English sounds, eliminating any influence from written representations or prior knowledge of word pronunciation. The experimental word list is: w_1:RALISAR, w_2:SHEEBINGS, w_3:BADUNLOT, w_4:MASIGAN, w_5:NAVIKLY, w_6:TAGAMAUGH, and w_7:HICKOMAY.

3.2 Data Recording

Participants were divided into two groups based on the pronunciation teaching strategy. The first group (RPI_G1) learned pronunciation through a repetition learning strategy, while the second group (RPI_G2) engaged in interactive recording sessions with an English teacher, utilising a repetition with feedback teaching strategy. For RPI_G1, the data recording process was conducted through a dedicated website. Participants had two options: online recording using their own setup while following provided instructions or participating in an in-person recording session at the university. For online recordings, participants were instructed to ensure a quiet environment without background noise and use a good-quality microphone. For in-person recordings, a meeting pod with sound isolation walls and headsets with built-in microphones was available. Participants listened to native speaker audio files, recorded their own pronunciation for each word, and were not allowed to replay the audio files or their own recordings during the session. RPI_G2 sessions took place at the university using a microphone positioned between the teacher and learner, who were seated approximately 50 cm apart and facing the same direction to prevent feedback from non-verbal cues. The microphone was placed 30 cm from each participant and 100 cm from the laptop running Audacity software (Audacity, 2017) for recording. Figure 1 illustrates the described recording setup. The teacher and learner were instructed to maintain a consistent volume level of around 60-70 decibels, speaking clearly and loudly enough to be easily understood without shouting. Recorded data were manually trimmed using Audacity software to include teacher pronunciation, learner pronunciation, and feedback. As described in Figure 2, in both RPI_G1 and RPI_G2, each of the 7 words was pronounced 12 times by each of the 25 learners during their individual recording sessions. In RPI_G1, one audio file was used as a reference for each word, recorded by a native teacher. In RPI_G2, learners repeated the words after the teacher, and feedback was provided. The RPI data comprises the final recordings from 50 L2 speakers and contributions from a native English teacher, resulting in a total of 6116 utterances with a total duration of 4 hours, 45 minutes, and 39 seconds.

4 Pronunciation Assessment Metrics

This section introduces a framework for evaluating the effectiveness of different automatic pronunciation assessment metrics in the context of L2 pronunciation learning. Various automatic pronunciation assessment metrics, denoted as Q_n, where n is the metric ID, $n \in 1, 2, 3, \ldots, 18$, are employed. The pronunciation score for a learner L repeating a word w_d for the i-th time, compared to the reference T, is represented by the notation $y_{Q_n,r,w_d,i}$. This score is calculated using the following equation:

$$y_{Q_n,r,w_d,i} = Q_n(L_{r,w_d,i}, T_{w_d,i}) \qquad (1)$$

where r is the learner ID, $r \in 1, 2, 3, \ldots, 50$, w_d is the word ID, $d \in 1, 2, 3, \ldots, 7$, and i is the repetition number, $i \in 1, 2, 3, \ldots, 12$. To illustrate the use of Equation (1), consider the following example: $L_{1,w_3,6}$ refers to the first learner repeating the third word from the word list in the sixth repetition. $T_{w_3,6}$ refers to the teacher repeating the third word from the word list in the sixth repetition. $y_{Q_1,1,w_3,6}$ represents the pronunciation score using metric Q_1 for the first learner and the third word in the sixth repetition. Table 1 summaries all the pronunciation assessment metrics for each Q_n.

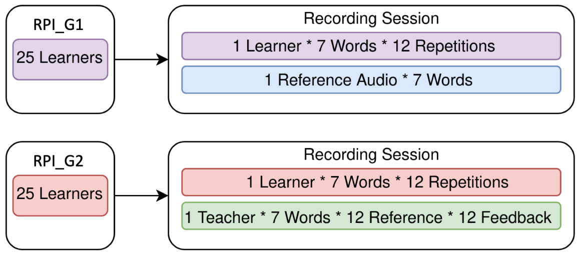

Figure 2: Description of the recording sessions for learners in both groups, (RPI_G1) using a repetition pronunciation teaching strategy, and (RPI_G2), using a repetition with feedback pronunciation teaching strategy.

4.1 Automatic Phone Recognition

Two distinct automatic phone recognises were tested to obtain phoneme sequences for both the native L1 teacher, serving as the reference, and the learner. The first recogniser, Allosaurus, is a universal phone recognition system trained with a multilingual allophone system (Li et al., 2020). The English models were trained on the VoxForge, Tedlium (Rousseau et al., 2012), and Switchboard (Godfrey et al., 1992) corpora. The PER for the recognised phonemes in relation to the reference phonemes serves as automatic pronunciation assessment metric Q_1. The second recogniser is a transformers-based model, a large-scale multilingual pre-trained model that uses the wav2vec 2.0 objective, as described in (Phy, 2022). In the context of speech recognition, XLS-R demonstrates significant improvements over recent models, achieving a relative error rate reduction of 20%-33% on average. This model is specifically trained on the TIMIT corpus (Garofolo et al., 1993), which includes speech recordings from 630 native speakers along with detailed phonetic transcriptions. The PER obtained with this model is denoted as Q_2. For example, consider the word "Ralisar". The recognised phoneme sequence using Q_2 for the learner is:

[lælɪsɑl]

The recognised phoneme sequence for the teacher is:

[rælɪsɑr]

In this case, the PER is 28.57%. Here's an example of evaluation using Q_2. For the learner with ID 2 and word ID 1, in the third repetition:

$$y_{Q_2,2,w_1,3} = Q_2(L_{2,w_1,3}, T_{w_1,3})$$

$$y_{Q_2,2,w_1,3} = 28.57$$

4.2 Self-Supervised Models

This section explores the use of self-supervised models for automatic pronunciation assessment by computing the acoustic representation-based distortion between learner and reference utterances. The distortion is calculated by aligning features extracted from the learner and reference audio using Dynamic Time Warping (DTW) and measuring the Euclidean distance between the aligned features. Several self-supervised models are employed in this study, including Wav2Vec 2.0 (Baevski et al., 2020), XLS-R (Babu et al., 2021), HuBERT (Hsu et al., 2021), and WavLM (Chen et al., 2022). These models are pre-trained on large amounts of unlabelled speech data and fine-tuned on labelled datasets to learn meaningful representations of speech at different linguistic levels. The experiments involve extracting features from different layers (layers 5, 12, 19, and the final layer) of the HuBERT, WavLM, XLS-R, and Wav2Vec 2.0 models. The notation for the automatic pronunciation assessment metrics based on these models is summarised in Table 1.

4.3 Goodness of Pronunciation (GOP)

The original GOP proposal aimed to derive a posterior per phoneme probability using a Gaussian Mixture Model-Hidden Markov Model (GMM-HMM). While in principle this is conceptually the right approach to assess pronunciations, it brings a range of problems. Using a Deep Neural Network-Hidden Markov Model (DNN-HMM)-based native acoustic model improves upon issues of estimation (Kim et al., 1997), but data-related drawbacks remain. Here, posterior probabilities for a set of senones are derived directly from a DNN, using alignments derived from the same model (Sudhakara et al., 2019). The acoustic model was trained on the LibriSpeech corpus (Panayotov et al., 2015), which is derived from LibriVox audiobooks and consists of about 1000 hours of read speech. The GOP model derived from here is further denoted with Q_{16}. A further GOP system using the same approach was trained on the WSJCAM0 corpus (Robinson et al., 1995), denoted with Q_{17}. WSJCAM0 contains read British English speech sentences. It was specifically designed for constructing and evaluating speaker-independent speech recognition systems in the early days of ASR development and has been used for GOP model training in different contexts. The corpus consists of recordings from 140 speakers, each delivering about 110 utterances. Finally, the GOP feature-based Transformer (GOPT) has been employed (Q_{18}) (Gong et al., 2022). The model is suggested to estimate pronunciation quality at multiple granularities and trained to predict the quality from multiple aspects using a transformer. First, an acoustic model is trained on LibriSpeech. The log phone posterior and the log posterior ratio between the canonical phone and the one with the highest score are used as GOP features. Then, the transformer takes the features to predict phoneme scores, word scores, and utterance-level scores.

5 Statistical Analysis

5.1 Normalisation

The metrics mentioned Table 1 all obtain values in different ranges. For comparability, it is desirable to have all scores in the same range. For this purpose, min-max normalisation is applied. Each value of $y_{Q_n,r,w_d,i}$ is calculated using Equation (1) and then normalised using Equation (2).

$$y'Qn, r, w_d, i = \frac{y_{Q_n,r,w_d,i} - y_{\min,Q_n}}{y_{\max,Q_n} - y_{\min,Q_n}} \qquad (2)$$

where $y'Qn, r, w_d, i$ is the normalised pronunciation score, $y_{\min,Q_n}$ is the minimum score among all pronunciation scores for metric Q_n. $y_{\max,Q_n}$ is the

Table 1: A list of assessment metrics with brief descriptions. The arrows represent the change in pronunciation score when the pronunciation improves. For example, the down arrow (↓) represents that a decrease in the score indicates improvement in pronunciation.

Q_n	**Metrics Description**
$Q_1\downarrow$	Allosaurus PER
$Q_2\downarrow$	Wav2vec 0.2-xls PER
$Q_3\downarrow$	HuBERT layer 5
$Q_4\downarrow$	WavLM layer 5
$Q_5\downarrow$	XLS-R layer 5
$Q_6\downarrow$	HuBERT layer 12
$Q_7\downarrow$	WavLM layer 12
$Q_8\downarrow$	XLS-R layer 12
$Q_9\downarrow$	HuBERT layer 19
$Q_{10}\downarrow$	WavLM layer 19
$Q_{11}\downarrow$	XLS-R layer 19
$Q_{12}\downarrow$	HuBERT layer 24
$Q_{13}\downarrow$	WavLM layer 24
$Q_{14}\downarrow$	XLS-R layer 24
$Q_{15}\downarrow$	Wav2Vec 2.0
$Q_{16}\uparrow$	GOP with LibriSpeech
$Q_{17}\uparrow$	GOP with WSJCAM0
$Q_{18}\uparrow$	GOPT

maximum score among all pronunciation scores for metric Q_n.

5.2 Pronunciation Learning Rate

To measure pronunciation improvement for each L2 learner r, the pronunciation learning rate for each learner r is computed using Equation (3). This is calculated by averaging the slopes of linear regression lines, each associated with a specific word. These slopes represent the rate of change in pronunciation scores, as determined by Q_n, with respect to the repetition number of each word.

$$P_{r,Q_n} = \frac{1}{7}\sum_{d=1}^{7}$$

$$\frac{\sum_{i=1}^{12}(x_i - \bar{x})(y'Qn,r,w_d,i - \overline{y'Qn,r,w_d,i})}{\sum_{i=1}^{12}(x_i - \bar{x})^2} \tag{3}$$

where $\bar{x}$ and $\overline{y'Qn,r,w_d,i}$ are the mean values for repetition number and the normalised pronunciation score, respectively. x is the repetition number $\in \{1,2,3,...,12\}$. and w_d represents the word ID, $d \in \{1,2,3,...,7\}$.

Negative slopes in automatic phone recognises denoted by Q_1 and Q_2, show a decrease in PER during repetition which should indicate learning progress. Negative slopes for self-supervised model metrics (Q_3 to Q_{15}) suggest a reduction in the distance between reference representations over the repetition period. Finally, a positive slope in the GOP metric, Q_{16} to Q_{18}, implies an increase in pronunciation quality during repetition.

6 Results

The results section presents the key findings of this study, focusing on two main aspects: the influence of feedback on L2 pronunciation learning and the impact of repetition on pronunciation scores. First, the pronunciation learning rates of two groups of learners (RPI_G1 and RPI_G2) are compared using the RPI corpus to assess the effectiveness of providing feedback during repetition. RPI_G1 engaged in a repetition-only learning strategy, while RPI_G2 received feedback during the repetition process. Second, the influence of repetition on pronunciation scores is examined by analysing the averaged pronunciation scores for each word repetition across various pronunciation assessment metrics. The relationship between the initial English proficiency level of L2 learners and their pronunciation skill was examined in Appendix A. Further analysis of word-level pronunciation improvement through repetition is provided in Appendix C.

6.1 Influence of Feedback on L2 Pronunciation Learning

In order to answer **RQ1**, the influence of feedback during repetition on the effectiveness of L2 pronunciation learning strategies is assessed by comparing the learning rates of two groups of learners in this section using the RPI corpus. The average pronunciation learning rate for RPI_G1 is calculated using Equation 4, and the average pronunciation learning rate for RPI_G2 is calculated using Equation 5.

$$P_{RPI\_G1,Q_n} = \frac{1}{25}\sum_{r=1}^{25} P_{r,Q_n} \tag{4}$$

where P_r is the pronunciation learning rate for each learner $r \in \{1,2,3,...,25\}$.

$$P_{RPI\_G2,Q_n} = \frac{1}{25}\sum_{r=26}^{50} P_{r,Q_n} \tag{5}$$

where P_r is the pronunciation learning rate for each learner $r \in \{26,27,28,...,50\}$.

Table 2 summaries the pronunciation learning rates of RPI_G1 and RPI_G2 across all pronunciation assessment metrics. The learning rates for

RPI_G2 indicate an improvement in pronunciation during repetition for all metrics except Q_1, while RPI_G1 shows the opposite trend, with the exception of Q_2. These findings support the hypothesis that the repetition with feedback strategy has a positive effect on L2 pronunciation learning. The positive learning rates for RPI_G2 across most metrics demonstrate the effectiveness of providing feedback to learners during the repetition process. Learners in RPI_G2 were able to incorporate the feedback to make significant improvements in their pronunciation over the repetitions. The consistency of this finding across multiple metrics strengthens the credibility of the results and highlights the robustness of the feedback-based approach. Conversely, the negative learning rates for RPI_G1 underscore the limitations of relying solely on repetition without feedback for pronunciation improvement. Learners in RPI_G1 may have struggled to perceive their own mistakes and make the necessary adjustments to enhance their pronunciation skills. The contrasting results between RPI_G1 and RPI_G2 emphasise the crucial role of feedback in the language learning process. Feedback provides learners with valuable information about their performance, enabling them to focus on specific areas that need improvement. These findings suggest that incorporating feedback into pronunciation training can substantially enhance learning outcomes, whereas relying exclusively on repetition may not yield the desired results. The influence of feedback on L2 pronunciation consistency and the influence of repetition on L2 pronunciation learning are examined in Appendix B.

6.2 Comparison of assessment metrics Q_n

This section delves into the findings related to **RQ2**, which focuses on the comparison of various pronunciation assessment metrics.

6.2.1 Q_n Consistency

The mean variance of each Q_n measures how far a set of scores is spread out from their average value. A lower variance indicates a more consistent metric across learners, while a higher variance suggests greater variability. The mean variance is computed by averaging the variance of all normalised pronunciation scores for all learners using Equation (6).

Table 2: The pronunciation learning rate of RPI_G1 and RPI_G2 across all assessment metrics Q_n.

Q_n	P_{RPI_G1,Q_n}	P_{RPI_G2,Q_n}
$Q_1 \downarrow$	5.14×10^{-6}	1.58×10^{-7}
$Q_2 \downarrow$	-1.67×10^{-7}	-2.41×10^{-7}
$Q_3 \downarrow$	5.28×10^{-5}	-3.65×10^{-5}
$Q_4 \downarrow$	4.38×10^{-5}	-1.38×10^{-5}
$Q_5 \downarrow$	3.96×10^{-5}	-7.44×10^{-6}
$Q_6 \downarrow$	7.57×10^{-5}	-1.06×10^{-4}
$Q_7 \downarrow$	8.42×10^{-5}	-6.49×10^{-5}
$Q_8 \downarrow$	4.82×10^{-5}	-6.93×10^{-5}
$Q_9 \downarrow$	9.56×10^{-5}	-2.18×10^{-4}
$Q_{10} \downarrow$	1.08×10^{-4}	-1.53×10^{-4}
$Q_{11} \downarrow$	7.82×10^{-5}	-2.04×10^{-4}
$Q_{12} \downarrow$	3.80×10^{-4}	-1.45×10^{-3}
$Q_{13} \downarrow$	4.02×10^{-4}	-7.46×10^{-4}
$Q_{14} \downarrow$	6.06×10^{-5}	-7.70×10^{-4}
$Q_{15} \downarrow$	7.81×10^{-7}	-1.22×10^{-7}
$Q_{16} \uparrow$	-6.9×10^{-8}	9.11×10^{-8}
$Q_{17} \uparrow$	-1.06×10^{-7}	2.47×10^{-7}
$Q_{18} \uparrow$	-2.42×10^{-8}	2.03×10^{-7}

The results are plotted in Figure

$$\sigma^2 Q_n = \frac{1}{50}\sum_{r=1}^{50}\frac{\sum_{w=1}^{7}(\sum_{i=1}^{12}(y'_{Q_n,r,w,i}-\mu))^2}{(W*I)-1} \tag{6}$$

where μ is the mean of all pronunciation scores for 50 learners, W is number of pseudo-words which is 7 and I is number of repetition which is 12.

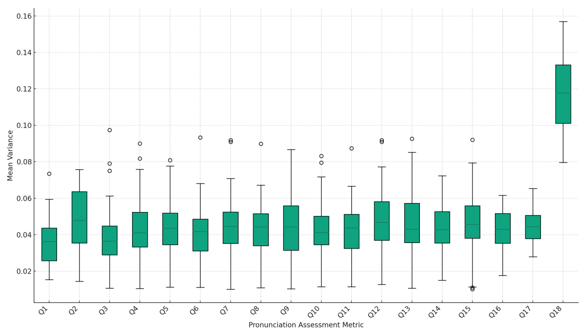

Figure 3: Mean variance of assessment metrics Q_n.

As Figure 3 shows, most Q_n have mean variances within a similar range. Metrics with lower mean variance, such as Q_1 and Q_3, are more consistent across different learners, suggesting more uniformity in their values. In contrast, Q_{18} has a noticeably higher mean variance compared to the others, indicating that its values vary more significantly among learners. The consistency of pronunciation assessment metrics is crucial for researchers and language educators. Metrics with lower mean variance provide more reliable and stable measurements of learners' performance, making it easier to compare progress across different individuals.

6.2.2 Correlation Between Q_n

The PCC between all pronunciation assessment metrics Q_n has been calculated and is shown in Figure 4. For each learner, the pronunciation learning rate P_{r,Q_n} is calculated using Equation (1) with a specific Q_a and denoted as P_{r,Q_a}, then calculated using another Q_b and denoted as P_{r,Q_b}. As

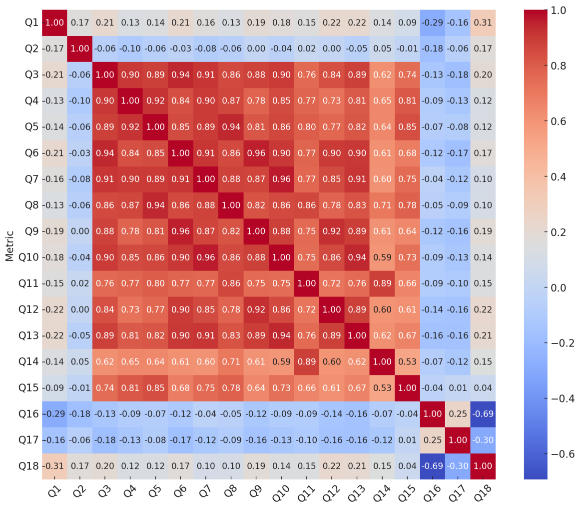

Figure 4: PCC between various assessment metrics Q_n.

Figure 4 shows, pairs of metrics with correlation coefficients close to 1 indicate a strong positive relationship. Q_3 and Q_6, Q_4 and Q_5, Q_5 and Q_8 show strong positive correlations over 0.9, and all of them are categorised as self-supervised models. This suggests that these self-supervised models capture similar aspects of pronunciation learning and provide consistent measurements of learners' progress. The high correlation among these metrics implies that they could potentially be used interchangeably or in combination to assess pronunciation development. Correlation coefficients that are positive but less than 0.5 indicate a moderate to weak positive relationship. Metrics like Q_1 with Q_3, Q_1 with Q_5, and Q_2 with Q_{18} fall into this category. The moderate to weak correlations between these metrics suggest that they capture different aspects of pronunciation and may provide complementary information about learners' performance. Q_1 and Q_2 are related to the same pronunciation assessment metrics category, which is automatic phone recognisers, while Q_{18} is GOPT. The weak correlation between the automatic phone recognisers and GOPT indicates that these metrics assess pronunciation from different perspectives and may offer distinct insights into learners' development.

7 Conclusion

This research provides valuable insights into the role of feedback and repetition in L2 English pronunciation learning for Chinese learners. The collection of the RPI corpus enabled a data-driven investigation comparing repetition with and without feedback. By employing a diverse set of automated pronunciation assessment metrics, the study presents a comprehensive evaluation of pronunciation improvement over multiple repetitions. The use of automated assessment methods is crucial in providing objective and reliable evaluations of pronunciation performance, overcoming the limitations of human evaluation, which often suffers from subjectivity and weak correlations among raters. The results demonstrate the positive impact of feedback on pronunciation learning rates, emphasising the importance of incorporating feedback into pronunciation training. The analysis of pronunciation assessment metrics reveals the consistency and correlation among different approaches, with self-supervised models showing promise in capturing pronunciation development. These findings have implications for language educators and researchers. Incorporating feedback into repetition-based pronunciation exercises can enhance learning outcomes. Furthermore, exploring multiple assessment metrics provides a more comprehensive understanding of learners' pronunciation progress. The study highlights the value of automated assessment in providing consistent and reliable measures of pronunciation performance.

8 Limitations

The current study has several limitations that should be acknowledged. Firstly, the RPI corpus is limited to Chinese learners of English, and the findings may not generalise to learners from other L1 backgrounds. Additionally, the study focused on a specific set of pseudo-words, and the effectiveness of the learning strategies and assessment metrics may vary with different word sets or authentic words. Furthermore, the long-term retention of pronunciation improvements was not investigated, and future research should explore the sustainability of learning gains. Moreover, the sample size of

50 learners, while sufficient for the current study, could be expanded in future research to increase the robustness of the findings. Lastly, the study did not control for individual differences in learners' aptitude, motivation, or prior pronunciation proficiency, which may influence their responsiveness to the learning strategies.

9 Preserving Anonymity and Ethics

Participants in this study were given a document called the Participant Information Sheet and Consent Forms, which had information in both English and Chinese to ensure clear understanding. These documents were approved by the Research Ethics Committee. Each participant received these documents one week before the recording session. The Participant Information Sheet contained details about the project, including why we specifically focused on Chinese speakers. It emphasised the voluntary nature of participation, allowing individuals to withdraw from the project at any time without providing a reason. Participants were encouraged to ask questions about the study after completing their participation. The information sheet outlined the steps participants would take, highlighted potential disadvantages and risks, and explained how the collected data would be utilised and stored. The university, acting as the data controller, assured secure and anonymous storage and transportation of the data. Anonymised data would be retained for at least 10 years after the study's conclusion, with ongoing reviews by the university to assess the necessity of continued retention. Contact details were also provided for any inquiries. It's important to note that the collected data remained anonymised, with no collection of names or gender information. Only age and IELTS results were gathered.

10 Acknowledgments

References

Chesta Agarwal and Pinaki Chakraborty. 2019. A review of tools and techniques for computer aided pronunciation training (capt) in english. *Education and Information Technologies*, 24:3731–3743.

Team Audacity. 2017. Audacity. *The Name Audacity (R) Is a Registered Trademark of Dominic Mazzoni Retrieved from http://audacity. sourceforge. net*.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al. 2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Alice YW Chan. 2007. The acquisition of english word-final consonants by cantonese esl learners in hong kong. *Canadian Journal of Linguistics/Revue canadienne de linguistique*, 52(3):231–253.

Xu Chang. 2021. Oral english in china. In *2021 5th International Seminar on Education, Management and Social Sciences (ISEMSS 2021)*, pages 575–579. Atlantis Press.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Zhengyang Chen, Shuai Wang, and Yanmin Qian. 2021. Self-supervised learning based domain adaptation for robust speaker verification. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5834–5838. IEEE.

Isabelle Darcy. 2018. Powerful and effective pronunciation instruction: How can we achieve it?. *Catesol Journal*, 30(1):13–45.

John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, and David S Pallett. 1993. Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1. *NASA STI/Recon technical report n*, 93:27403.

Abbas Pourhosein Gilakjani. 2012. A study of factors affecting efl learners' english pronunciation learning and the strategies for instruction. *International journal of humanities and social science*, 2(3):119–128.

Abbas Pourhossein Gilakjani and Mohammad Reza Ahmadi. 2011. Why is pronunciation so difficult to learn?. *English language teaching*, 4(3):74–83.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, speech, and signal processing, ieee international conference on*, volume 1, pages 517–520. IEEE Computer Society.

Yuan Gong, Ziyi Chen, Iek-Heng Chu, Peng Chang, and James Glass. 2022. Transformer-based multi-aspect multi-granularity non-native english speaker pronunciation assessment. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7262–7266. IEEE.

Feifei Han. 2013. Pronunciation problems of chinese learners of english. *Ortesol Journal*, 30:26–30.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Jingtao Hu, Lei Gao, Xiaoping Bai, Taochang Li, and Xiaoguang Liu. 2015. Review of research on automatic guidance of agricultural vehicles. *Transactions of the Chinese Society of Agricultural Engineering*, 31(10):1–10.

Elaf Islam, Chanho Park, and Thomas Hain. 2023. Exploring speech representations for proficiency assessment in language learning. In *9th Workshop on Speech and Language Technology in Education (SLaTE) Proceedings*, pages 151–155. International Speech Communication Association (ISCA).

Syed Mazharul Islam. 2020. Segmental errors in english pronunciation of non-native english speakers. *Journal of Education and Social Sciences*, 16(1):14–24.

Sandra Kanters, Catia Cucchiarini, and Helmer Strik. 2009. The goodness of pronunciation algorithm: a detailed performance study.

Emmanuel Keuleers and Marc Brysbaert. 2011. Detecting inherent bias in lexical decision experiments with the ld1nn algorithm. *The Mental Lexicon*, 6(1):34–52.

Subash Khanal, Michael T Johnson, Mohammad Soleymanpour, and Narjes Bozorg. 2021. Mispronunciation detection and diagnosis for mandarin accented english speech. In *2021 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 62–67. IEEE.

Yassine Kheir, Ahmed Ali, and Shammur Chowdhury. 2023. Automatic pronunciation assessment-a review. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8304–8324.

Eesung Kim, Jae-Jin Jeon, Hyeji Seo, and Hoon Kim. 2022. Automatic pronunciation assessment using self-supervised speech representation learning. *arXiv preprint arXiv:2204.03863*.

Yoon Kim, Horacio Franco, and Leonardo Neumeyer. 1997. Automatic pronunciation scoring of specific phone segments for language instruction. In *Fifth European Conference on Speech Communication and Technology*.

Daniel Korzekwa, Jaime Lorenzo-Trueba, Thomas Drugman, and Bozena Kostek. 2022. Computer-assisted pronunciation training—speech synthesis is almost all you need. *Speech Communication*, 142:22–33.

Takuya Kunihara, Chuanbo Zhu, Nobuaki Minematsu, and Noriko Nakanishi. 2022. Gradual improvements ob-served in learners' perception and production of l2 sounds through continuing shadowing practices on a daily basis. In *Proc. Interspeech*.

Diane Larsen-Freeman. 2012. On the roles of repetition in language teaching and learning. *Applied Linguistics Review*, 3(2):195–210.

Juyeon Lee. 2013. The effects of discourse types on the use of english articles by korean learners of english: Oral vs. written narratives. *English Language Teaching*, 6(8):33–43.

Xinjian Li, Siddharth Dalmia, Juncheng Li, Matthew Lee, Patrick Littell, Jiali Yao, Antonios Anastasopoulos, David R Mortensen, Graham Neubig, Alan W Black, et al. 2020. Universal phone recognition with a multilingual allophone system. In *ICASSP*, pages 8249–8253. IEEE.

Beate Luo. 2016. Evaluating a computer-assisted pronunciation training (capt) technique for efficient classroom instruction. *Computer assisted language learning*, 29(3):451–476.

Jan Malucha. 2022. Software tool for pronunciation training of specific english terminology. In *2022 New Trends in Signal Processing (NTSP)*, pages 1–5. IEEE.

Wolfgang Menzel, Eric Atwell, Patrizia Bonaventura, Daniel Herron, Peter Howarth, Rachel Morton, and Clive Souter. 2000. The isle corpus of non-native spoken english. In *Proceedings of LREC 2000: Language Resources and Evaluation Conference, vol. 2*, pages 957–964. European Language Resources Association.

Edmilson Morais, Ron Hoory, Weizhong Zhu, Itai Gat, Matheus Damasceno, and Hagai Aronowitz. 2022. Speech emotion recognition using self-supervised features. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6922–6926. IEEE.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, Speech and Signal Processing*, pages 5206–5210. IEEE.

Vitou Phy. 2022. Automatic Phoneme Recognition on TIMIT Dataset with Wav2Vec 2.0. If you use this model, please cite it using these metadata.

Mirco Ravanelli, Jianyuan Zhong, Santiago Pascual, Pawel Swietojanski, Joao Monteiro, Jan Trmal, and Yoshua Bengio. 2020. Multi-task self-supervised learning for robust speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6989–6993. IEEE.

Monica Richards. 2011. Helping chinese learners distinguish english/l/and/n. *Pronunciation in Second Language Learning and Teaching Proceedings*, 3(1).

Tony Robinson, Jeroen Fransen, David Pye, Jonathan Foote, and Steve Renals. 1995. Wsjcamo: a british english speech corpus for large vocabulary continuous speech recognition. In *1995 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 81–84. IEEE.

Pamela M Rogerson-Revell. 2021. Computer-assisted pronunciation training (capt): Current issues and future directions. *Relc Journal*, 52(1):189–205.

Anthony Rousseau, Paul Deléglise, and Yannick Esteve. 2012. Ted-lium: an automatic speech recognition dedicated corpus. In *LREC*, pages 125–129.

Hooman Saeli, Payam Rahmati, and Mohammadreza Dalman. 2021. Oral corrective feedback on pronunciation errors: The mediating effects of learners' engagement with feedback. *Advances in Language and Literary Studies*, 12(4):68–78.

Kazuya Saito and Roy Lyster. 2012. Effects of form-focused instruction and corrective feedback on l2 pronunciation development of r by japanese learners of english. *Language learning*, 62(2):595–633.

Kavita Sheoran, Arpit Bajgoti, Rishik Gupta, Nishtha Jatana, Geetika Dhand, Charu Gupta, Pankaj Dadheech, Umar Yahya, and Nagender Aneja. 2023. Pronunciation scoring with goodness of pronunciation and dynamic time warping. *IEEE Access*, 11:15485–15495.

Sweekar Sudhakara, Manoj Kumar Ramanathi, Chiranjeevi Yarra, and Prasanta Kumar Ghosh. 2019. An improved goodness of pronunciation (gop) measure for pronunciation evaluation with dnn-hmm system considering hmm transition probabilities. In *INTERSPEECH*, volume 2, pages 954–958.

Cristian Tejedor-García, David Escudero-Mancebo, Enrique Cámara-Arenas, César González-Ferreras, and Valentín Cardeñoso-Payo. 2020. Assessing pronunciation improvement in students of english using a controlled computer-assisted pronunciation tool. *IEEE Transactions on Learning Technologies*, 13(2):269–282.

Andros Tjandra, Diptanu Gon Choudhury, Frank Zhang, Kritika Singh, Alexis Conneau, Alexei Baevski, Assaf Sela, Yatharth Saraf, and Michael Auli. 2022. Improved language identification through cross-lingual self-supervised learning. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6877–6881. IEEE.

Xinchun Wang and Jidong Chen. 2020. The acquisition of mandarin consonants by english learners: The relationship between perception and production. *Languages*, 5(2):20.

Yu Wang, Mark JF Gales, Kate M Knill, Konstantinos Kyriakopoulos, Andrey Malinin, Rogier C van Dalen, and Mohammad Rashid. 2018. Towards automatic assessment of spontaneous spoken english. *Speech Communication*, 104:47–56.

Eun Jung Yeo, Hyungshin Ryu, Jooyoung Lee, Sunhee Kim, and Minhwa Chung. 2023. Comparison of l2 korean pronunciation error patterns from five l1 backgrounds by using automatic phonetic transcription. *arXiv preprint arXiv:2306.10821*.

Feng Zhang, Chao Huang, Frank K Soong, Min Chu, and Renhua Wang. 2008. Automatic mispronunciation detection for mandarin. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5077–5080. IEEE.

Junbo Zhang, Zhiwen Zhang, Yongqing Wang, Zhiyong Yan, Qiong Song, Yukai Huang, Ke Li, Daniel Povey, and Yujun Wang. 2021. speechocean762: An open-source non-native english speech corpus for pronunciation assessment. *arXiv preprint arXiv:2104.01378*.

Yanyan Zhang and Jing Xiao. 2014. An analysis of chinese students' perception and production of paired english fricatives: From an elf perspective. *Journal of Pan-Pacific Association of Applied Linguistics*, 18(1):171–192.

A Appendix A

A.1 Initial English Proficiency and L2 Pronunciation Skill

Before analysing the gradual improvement in L2 pronunciation learning, the relationship between the initial English proficiency level of L2 learners and their pronunciation skill was examined. To assess the learners' proficiency, recent IELTS scores were collected from each participant. The average pronunciation score and learning rate per learner were then calculated using Equation (1) and Equation (3), respectively. The PCC between the IELTS score and each average was computed to determine the strength and direction of the relationship. Table 3 presents the PCC values between the averaged pronunciation learning rate per learner and the IELTS score, as well as the PCC values between the averaged pronunciation score and the IELTS score for selected pronunciation assessment metrics (Q_n). As shown in Table 3, the correlation coefficients range from -0.16 to 0.13, indicating very weak relationships between the IELTS scores and both the average pronunciation learning rate and the average pronunciation score. Some of the correlations are even in the opposite direction, suggesting that higher IELTS scores do not necessarily correspond to better pronunciation skills or faster learning rates. These findings raise questions about the suitability of using IELTS scores as a predictor of L2 pronunciation proficiency. While IELTS is a widely recognised English language proficiency

Table 3: PCC between the averaged pronunciation learning rate per learner and the IELTS score, and PCC between the averaged pronunciation score and the IELTS score.

Q_n	PCC with Average Pronunciation Learning Rate	PCC with Average Pronunciation Score
Q_2	-0.12	0.06
Q_4	0.04	0.01
$Q6$	0.08	0.02
$Q11$	-0.01	0.13
$Q15$	0.02	0.05
$Q17$	-0.16	0.05

test, it may not provide a comprehensive assessment of pronunciation skills specifically. The weak correlations observed in this study suggest that alternative English pre-tests targeting pronunciation more directly may be needed to better understand the relationship between initial proficiency and pronunciation learning outcomes.

B Appendix B

B.0.1 Influence of Feedback on L2 Pronunciation Consistency

In this context, L2 pronunciation consistency refers to the extent to which learners in each group demonstrate stable and uniform pronunciation patterns across multiple repetitions. A stable pronunciation pattern means that learners maintain a consistent level of pronunciation accuracy throughout the repetitions, without significant variations or deviations. Pronunciation consistency can be inferred by examining the PCC between different pronunciation scores among learners in the same group. A higher correlation indicates a higher level of consistency, suggesting that learners in the same group exhibit similar pronunciation patterns across repetitions. Figures 5 display the PCC values between different pronunciation scores among learners in RPI_G1 and RPI_G2, respectively. As seen in the right figure, learners in RPI_G2 demonstrate higher correlation coefficients, indicating greater consistency in pronunciation compared to learners in RPI_G1. This observation underscores the significance of feedback in L2 pronunciation learning, as it suggests that providing feedback helps learners maintain a more consistent pronunciation pattern throughout the repetitions.

B.0.2 Influence of Repetition on L2 Pronunciation Learning

This section explores the impact of repetition by calculating REP_{Q_n}, which represents the averaged pronunciation scores for each word repetition, obtained using Equation (7).

$$REP_{Q_n} = \frac{1}{7}\sum_{w=1}^{7}\frac{1}{12}\sum_{i=1}^{12}\frac{1}{50}\sum_{r=1}^{50} y'Qn,r,w,i \tag{7}$$

Using Q_{12} as a pronunciation assessment metric, Figure 6 illustrates the averaged pronunciation scores for each repetition per word for RPI_G1 and RPI_G2. Repetition 2 shows the smallest averaged pronunciation score in RPI_G1, while Repetition 6 shows the smallest averaged pronunciation score in RPI_G2. These findings suggest that repeating the words about six times may lead to an improvement in pronunciation, as evidenced by the lower pronunciation scores at these repetition numbers. Although both scores fluctuate, the score for the RPI_G2 group has a downward tendency and appears to converge as repetition increases, indicating an overall improvement in pronunciation. In contrast, the curve for RPI_G1 shows an overall rise, suggesting a lack of consistent improvement in pronunciation without feedback. Using Equation (7) for all Q_n, Figure 7 indicates the repetition number at which the best pronunciation score occurs for each of RPI_G1 and RPI_G2. Here, the best pronunciation score refers to the smallest score among Q_1 to Q_{15} and the largest score among Q_{16} to Q_{18}. Repetition 2 is the point of best pronunciation in RPI_G1. In RPI_G2, Repetition 6 holds the position of best pronunciation, with Repetition 3 also being a notable point.

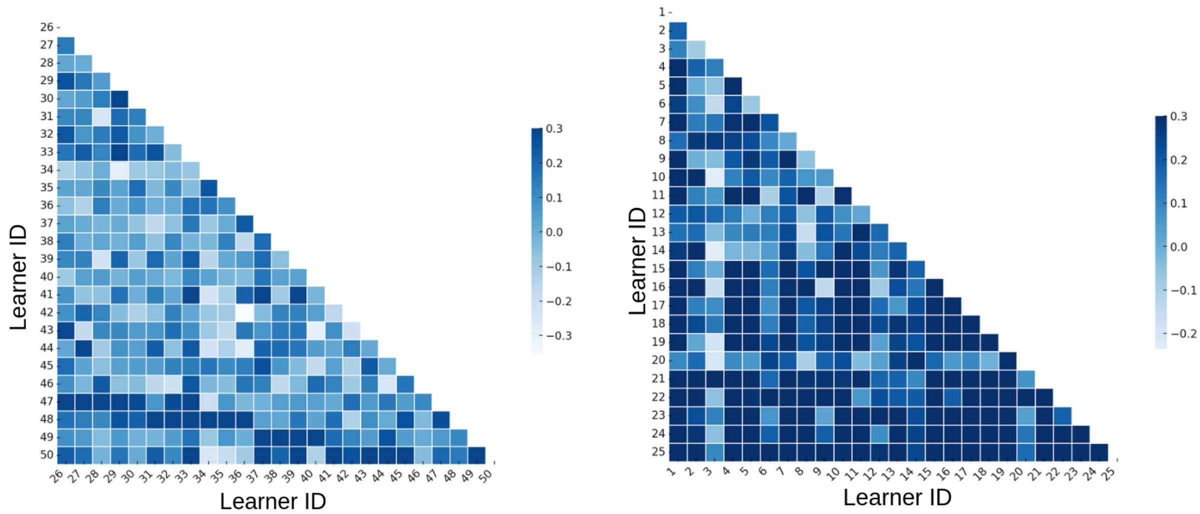

Figure 5: PCC between different pronunciation scores among learners in RPI_G1 (left) and PCC between different pronunciation scores among learners in RPI_G2 (right).

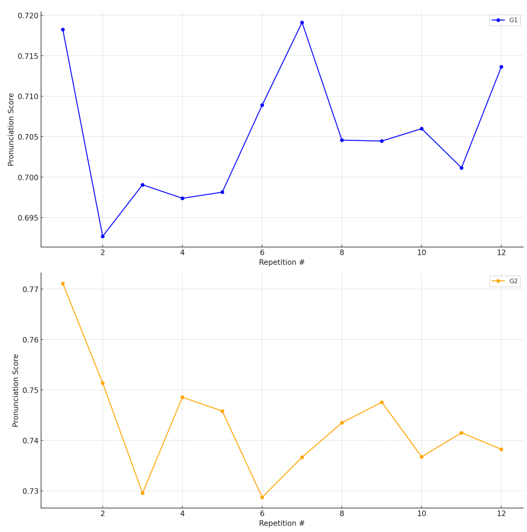

Figure 6: Averaged pronunciation scores for words per repetition using Q_{12} for RPI_G1 and RPI_G2.

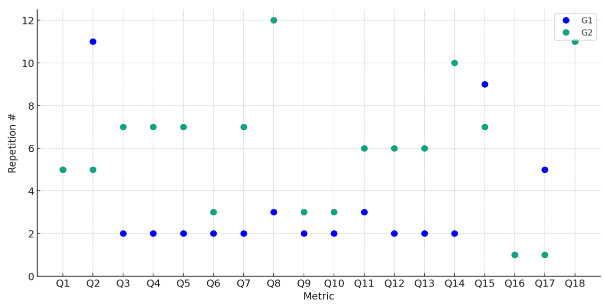

Figure 7: The repetition number at which the best pronunciation score occurs for each of RPI_G1 and RPI_G2 in all Q_n.

C Appendix C

C.1 Improvement in Word Pronunciation through L2 Learning

This section addresses the identification of words that exhibit pronunciation improvement with repetition. Based on Figure 6, the sixth repetition shows the best pronunciation score for RPI_G2, with a lower score compared to the first repetition, indicating better word pronunciation. Conversely, a higher score in the sixth repetition would suggest that the word is difficult to learn. For the assessment, Q_1, Q_3, and Q_{16} are selected based on their consistency and correlation, as discussed in Section 6.2. Figures 8, 9, and 10 illustrate the score changes for each of the seven pseudo-words using Q_1, Q_3, and Q_{16}, respectively.

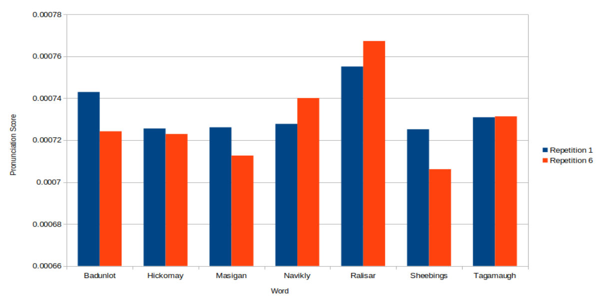

Figure 8: Pronunciation score for each of the seven alien words for RPI_G2 using $Q1 \downarrow$.

The pronunciation scores for **Badunlot**, **Masigan**, **Ralisar**, and **Sheebings** change consistently across the three metrics. The scores using Q_1 and Q_3 decrease, while those using Q_{16} increase, in-

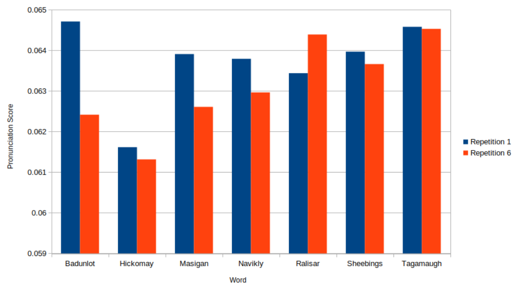

Figure 9: Pronunciation scores for each of the seven pseudo-words in RPI_G2 using $Q_3 \downarrow$.

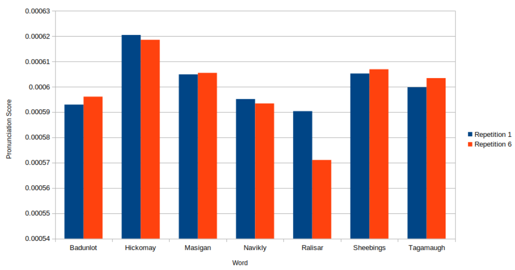

Figure 10: Pronunciation scores for each of the seven pseudo-words in RPI_G2 using $Q_{16} \uparrow$.

dicating an improvement in pronunciation. However, the score changes for **Hickomay**, **Navikly**, and **Tagamaugh** are inconsistent among the metrics. For example, the Q_1 score for **Navikly** increases, while the scores for Q_3 and Q_{16} decrease. In cases where the scores from the three metrics show inconsistent results, the decision regarding pronunciation improvement is made based on the majority. For instance, in the example above, the results for **Navikly** can be interpreted as a degradation in pronunciation, as indicated by Q_1 and Q_{16}. Similarly, the figures show that the pronunciation of five out of the seven words improves: **Badunlot**, **Hickomay**, **Masigan**, **Sheebings**, and **Tagamaugh**. In summary, this section demonstrates the use of multiple pronunciation assessment metrics to identify words that show improvement in pronunciation through repetition. By comparing the scores from the first and sixth repetitions, and considering the consistency of score changes across different metrics, it is possible to determine which words benefit from repetition in terms of pronunciation improvement.